ViPOcc: Leveraging Visual Priors from Vision Foundation Models for Single-View 3D Occupancy Prediction

Yi Feng[1], Yu Han[2], Xijing Zhang[1], Tanghui Li[1], Yanting Zhang[2], Rui Fan[1,3,4]*

[1]College of Electronics and Information Engineering, Tongji University
[2]School of Computer Science and Technology, Donghua University
[3]Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University
[4]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University
fengyi@ieee.org, 2232816@mail.dhu.edu.cn, {xijingzhang,thli}@tongji.edu.cn, ytzhang@dhu.edu.cn, rui.fan@ieee.org

Abstract

Inferring the 3D structure of a scene from a single image is an ill-posed and challenging problem in the field of vision-centric autonomous driving. Existing methods usually employ neural radiance fields to produce voxelized 3D occupancy, lacking instance-level semantic reasoning and temporal photometric consistency. In this paper, we propose ViPOcc, which leverages the visual priors from vision foundation models (VFMs) for fine-grained 3D occupancy prediction. Unlike previous works that solely employ volume rendering for RGB and depth image reconstruction, we introduce a metric depth estimation branch, in which an inverse depth alignment module is proposed to bridge the domain gap in depth distribution between VFM predictions and the ground truth. The recovered metric depth is then utilized in temporal photometric alignment and spatial geometric alignment to ensure accurate and consistent 3D occupancy prediction. Additionally, we also propose a semantic-guided non-overlapping Gaussian mixture sampler for efficient, instance-aware ray sampling, which addresses the redundant and imbalanced sampling issue that still exists in previous state-of-the-art methods. Extensive experiments demonstrate the superior performance of ViPOcc in both 3D occupancy prediction and depth estimation tasks on diverse public datasets.

Code — https://mias.group/ViPOcc

Introduction

As a key ingredient of environmental perception in autonomous driving, 3D occupancy prediction has garnered considerable attention in recent years (Wei et al. 2023; Huang et al. 2024; Tian et al. 2024). Early efforts tackle this problem through supervised learning, which requires extensive 3D human-labeled annotations and depth ground truth acquired using additional range sensors (Huang et al. 2023). More recently, neural radiance field (NeRF)-based approaches have emerged as promising techniques for unsupervised single-view 3D occupancy prediction (Wimbauer et al. 2023; Li et al. 2024), noted for their capability to render photorealistic images from novel viewpoints.

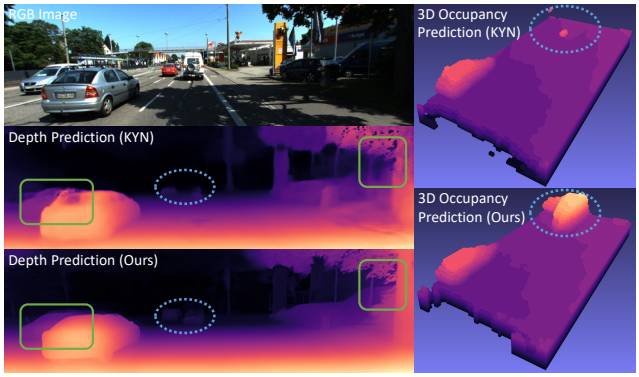

Figure 1: **Single-view 3D scene reconstruction results.** KYN (Li et al. 2024) struggles to recover clear object boundaries (green boxes) and exhibits poor reconstruction performance for distant objects (blue circles). ViPOcc outperforms KYN in both monocular depth estimation and 3D occupancy prediction tasks.

As a pioneering work, BTS (Wimbauer et al. 2023) estimates a 3D density field from a single view, relying solely on photometric consistency constraints across multiple views during training. Subsequent studies (Han et al. 2024; Li et al. 2024) adopt the same training strategy for 3D scene reconstruction but often underexploit temporal photometric and geometric constraints, resulting in inconsistent 3D occupancy predictions across adjacent frames.

Another growing trend is to unleash the potential of vision foundation models (VFMs) for comprehensive 3D scene representation. As a notable example, KYN (Li et al. 2024) leverages a large vision-language model to enrich 3D features with semantic information. However, as illustrated in Fig. 1, challenges remain, particularly with the frequent omission of critical instances, due to the indiscriminate random ray sampling process. SC-DepthV3 (Sun et al. 2024) uses predictions from LeReS (Yin et al. 2021) as pseudo depth for robust unsupervised depth estimation. Nevertheless, current VFMs generally produce monocular depth predictions with inherent scale ambiguity (Yang et al. 2024), which are not directly applicable to temporal photometric alignment. The recently proposed VFM, Depth Anything V2

*Corresponding author: Rui Fan.

(L. Yang *et al.* 2024), demonstrates exceptional zero-shot performance in metric depth estimation with fine-grained details. However, it experiences a significant performance decline due to domain discrepancies between the training and test data.

To address the aforementioned challenges, we introduce ViPOcc, a novel approach that leverages visual priors from VFMs for fine-grained, instance-aware 3D scene reconstruction. Unlike previous state-of-the-art (SoTA) methods that solely utilize photometric discrepancies as supervisory signals, our method incorporates a depth prediction branch, which fully exploits inter-frame photometric consistency and intra-frame geometric reconstruction consistency, enabling self-supervised training with spatial-temporal consistency constraints.

Specifically, we design an inverse depth alignment module that mitigates the discrepancies between VFM predictions and depth ground truth, leading to compelling metric depth estimation results. To further enhance both the efficiency and accuracy of 3D occupancy prediction, we develop a semantic-guided, non-overlapping Gaussian mixture (SNOG) sampler, which effectively addresses issues such as redundant ray sampling and the overlooking of crucial instances prevalent in previous methods. Additionally, we propose a temporal alignment loss and a reconstruction consistency loss, which further improve the quality of both metric depth and 3D occupancy predictions. Extensive experiments on the KITTI-360 and KITTI Raw datasets validate the effectiveness of each developed component and further demonstrate ViPOcc's superior performance over all existing SoTA methods.

In a nutshell, we present the following key contributions:

1. We propose **ViPOcc**, a single-view 3D **Occ**upancy prediction framework that incorporates **Vi**sual **P**riors from VFMs, achieving SoTA performance in both monocular depth estimation and 3d occupancy prediction tasks.
2. We introduce an inverse depth alignment module that effectively recovers the scale of the VFM's depth predictions while preserving their local visual details.
3. We present a SNOG sampler that guides the framework to focus more on crucial instances and avoid overlapping patches during ray sampling.
4. We establish a novel training paradigm that couples the unsupervised training of 3D occupancy prediction and monocular depth estimation using the proposed temporal alignment and reconstruction consistency losses.

Related Work

Single-View 3D Occupancy Prediction

Deriving voxelized 3D occupancy of a scene from a single image is a promising technique for achieving fine-grained geometric representation and comprehensive environmental understanding in 3D space (Zhang et al. 2024). As a pioneering work, MonoScene (Cao and De Charette 2022) leverages voxel features generated through view projection for occupancy regression. However, this method is not suitable for real-time multi-view 3D reconstruction due to the inefficiency of voxel representations. TPVFormer (Huang et al. 2023) extends it to a multi-camera setup by incorporating tri-perspective view representations. Despite their compelling performance, these supervised methods necessitate data with 3D ground truth, which requires labor-intensive human annotation. Recently, the study (Wimbauer et al. 2023) introduced BTS, a fully unsupervised method that uses perspective and fisheye video sequences to reconstruct driving scenes with NeRF-based volume rendering techniques. Following this work, KYN (Li et al. 2024) leverages meaningful semantic and spatial context for fine-grained 3D scene reconstruction. MVBTS (Han et al. 2024) combines density fields from multi-view images through knowledge distillation, achieving SoTA performance in handling occluded regions. Different from existing NeRF-based frameworks, we incorporate an additional depth prediction branch for spatial-temporal 3D occupancy alignment.

Visual Priors for 3D Scene Reconstruction

Previous studies (Li et al. 2024; Zhang et al. 2023a) have integrated visual priors from pre-trained VFMs into depth estimation and NeRF-based 3D scene reconstruction frameworks. Existing depth estimation methods typically utilize pre-inferred semantics for fine-grained feature representation and fusion (Guizilini et al. 2020b; Jung et al. 2021; Chen et al. 2023). Other studies (Kerr et al. 2023; Peng et al. 2023) leverage 2D visual priors for 3D feature representation and registration. KYN (Li et al. 2024) incorporates a pre-trained vision-language network for robust 3D feature representation, significantly improving 3D shape recovery. MonoOcc (Zheng et al. 2024) employs a pre-trained InternImage-XL (Wang et al. 2023) as its backbone for visual feature extraction and distillation. OccNeRF (Zhang et al. 2023a) utilizes frozen VFMs for 2D semantic supervision but faces challenges in detecting small instances due to the limitations of open-vocabulary models in capturing fine details. While these methods have successfully leveraged the strengths of VFMs for feature extraction, the informative visual priors from VFMs remain underutilized. In this paper, we leverages semantic priors from Grounded-SAM (Ren et al. 2024) and spatial priors from Depth Anything V2 (L. Yang *et al.* 2024) for efficient ray sampling and spatial-temporal 3D occupancy alignment.

Unsupervised Monocular Depth Estimation

Existing frameworks typically maximize photometric consistency across video sequences or stereo image pairs to estimate scale-invariant depth maps. SfMLearner (Zhou et al. 2017), the first reported study in this field, jointly estimates depth maps and camera poses between successive video frames by minimizing a photometric reprojection loss. Building on this method, Monodepth2 (Godard et al. 2019) introduces a minimum reprojection loss to address occlusion issues and an automasking loss to exclude moving objects that appear stationary relative to the camera. Subsequent studies mainly explored various network architectures (Wang et al. 2024; Watson et al. 2021), dynamic object filtering strategies (Sun and Hariharan 2024; Yin and Shi 2018), and additional constraints (Guizilini et al. 2020b;

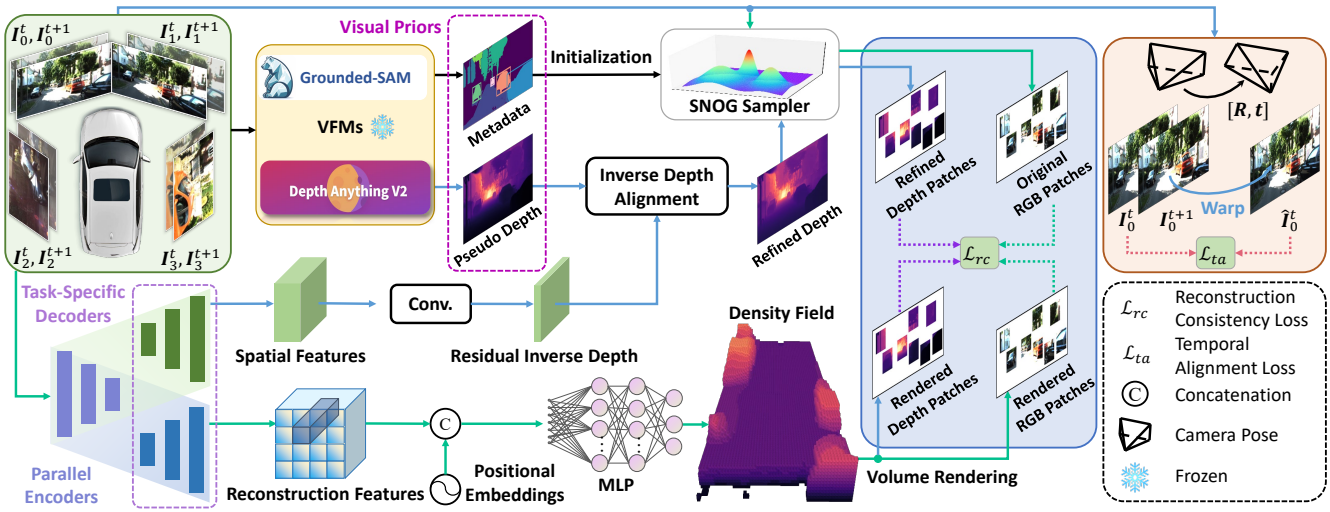

Figure 2: An illustration of our proposed **ViPOcc** framework. Unlike previous approaches that rely solely on NeRF for 3D scene reconstruction, ViPOcc introduces an additional depth prediction branch and an instance-aware SNOG sampler for temporal photometric alignment and spatial geometric alignment.

Schmied et al. 2023). Other NeRF-based frameworks (Wimbauer et al. 2023; Han et al. 2024) estimate metric depth maps through discrete volume rendering. However, their predictions often lack accuracy and fail to preserve clear object contours. In contrast, our proposed ViPOcc utilizes visual priors from VFMs to enable instance-aware ray sampling and fine-grained metric depth estimation.

Methodology

Problem Setup

Given an input RGB image $\boldsymbol{I}$ and its corresponding intrinsic matrix $\boldsymbol{K}$, we aim to reconstruct the 3D geometry of the entire scene with the voxelized density:

$$\sigma_{\boldsymbol{p}} = \mathcal{R}(\boldsymbol{p}, \boldsymbol{I}, \boldsymbol{K}, \boldsymbol{\Theta}), \tag{1}$$

where $\boldsymbol{p}$ denotes a 3D point in the reconstructed scene, and $\mathcal{R}(\cdot)$ represents the neural radiance field with learnable parameters $\boldsymbol{\Theta}$. $\sigma_{\boldsymbol{p}}$ can be further employed to produce a rendered RGB image $\hat{\boldsymbol{I}}_r$ and a rendered distance map $\hat{\boldsymbol{D}}_r$ using the following expressions:

$$\hat{\boldsymbol{I}}_r(\boldsymbol{p}_i) = \sum_{i=1}^{M} T_i \alpha_i c_{\boldsymbol{p}_i}, \quad \hat{\boldsymbol{D}}_r(\boldsymbol{p}_i) = \sum_{i=1}^{M} T_i \alpha_i d_i, \tag{2}$$

where $\alpha_i = 1 - \exp\left(-\sigma_{\boldsymbol{p}_i} ||\boldsymbol{p}_{i+1} - \boldsymbol{p}_i||_2\right)$ denotes the probability that the ray ends between $\boldsymbol{p}_i$ and $\boldsymbol{p}_{i+1}$, $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$ represents the accumulated transmittance, $c_{\boldsymbol{p}_i}$ denotes the sampled RGB value from other viewpoints, and d_i represents the distance between $\boldsymbol{p}_i$ and the ray origin.

Architecture Overview

As illustrated in Fig. 2, ViPOcc takes stereo image pairs $\boldsymbol{I}_{0,1}^{t,t+1}$ and rectified fisheye images $\boldsymbol{I}_{2,3}^{t,t+1}$ captured at timestamps t and $t+1$ as input. $\boldsymbol{I}_0^t$ is regarded as the principal frame, from which spatial features $\boldsymbol{F}_s$ and reconstruction features $\boldsymbol{F}_r$ are extracted using parallel encoders and task-specific decoders. During training, ViPOcc simultaneously generates 2D depth maps and 3D density fields from two separate branches. In the depth estimation branch, an inverse depth alignment module is designed to mitigate the domain discrepancy between depth priors from a VFM and the depth ground truth. The refined depth maps $\hat{\boldsymbol{D}}$ and the corresponding RGB images are then fed into our developed SNOG sampler for efficient ray sampling, producing instance-aware and non-overlapping patches. On the other hand, in the 3D occupancy prediction branch, $\boldsymbol{F}_r$ combined with positional embeddings $\boldsymbol{F}_p$ is passed through an MLP to predict a 3D density field, which is then utilized in volume rendering to generate depth and RGB patches. By enforcing reconstruction consistency across sampled RGB and depth patches, as well as temporal photometric consistency between adjacent principal frames, we achieve improved performance in both 3D occupancy prediction and metric depth estimation.

Inverse Depth Alignment

Unlike prior arts (Wimbauer et al. 2023; Han et al. 2024) that rely solely on NeRF-based reconstruction consistency to supervise framework training, we incorporate inter-frame photometric consistency and depth rendering consistency through a VFM-driven depth estimation branch. Pseudo depth maps $\boldsymbol{D}_p$ are first obtained from off-the-shelf VFMs like Depth Anything V2 (L. Yang *et al.* 2024). Nevertheless, as demonstrated in our experiments, the residuals between pseudo and ground-truth depth data exhibit dramatically deviated distributions. These deviations arise from significant domain gaps between real-world scenarios and the data on which VFMs are initially trained. Therefore, it is imperative to refine depth before utilizing it to introduce additional constraints for temporal photometric alignment. As discussed in

(He et al. 2016), neural networks often struggle to converge or maintain accuracy when fitting large ranges of numerical variations. It is thus plausible to fit residual inverse depth in our task, expressed as $\mathcal{F}(\boldsymbol{x}) := \frac{1}{\hat{\boldsymbol{D}}(\boldsymbol{x})} - \frac{1}{\boldsymbol{D}_p(\boldsymbol{x})}$, where $\mathcal{F}(\cdot)$ denotes the residual inverse depth function, $\hat{\boldsymbol{D}}$ represents the refined depth map, and $\boldsymbol{x}$ denotes a given 2D pixel. This function can be effectively fitted using spatial features $\boldsymbol{F}_s$ by formulating it as $\mathcal{F}(\boldsymbol{x}) = f(\boldsymbol{F}_s, \boldsymbol{\theta})$, where $f(\cdot)$ denotes a convolutional layer with learnable parameters $\boldsymbol{\theta}$. The refined depth map can therefore be yielded as follows:

$$\hat{\boldsymbol{D}}(\boldsymbol{x}) = \frac{1}{\frac{1}{\boldsymbol{D}_p(\boldsymbol{x})} + f(\boldsymbol{F}_s, \boldsymbol{\theta}) + \epsilon}, \tag{3}$$

where ϵ is a small constant used to prevent the denominator from being zero. $\hat{\boldsymbol{D}}$ can then be used to ensure inter-frame photometric consistency and depth rendering consistency.

Semantic-Guided Non-Overlapping Gaussian Mixture Sampler

Focusing on individual instances rather than the entire scene can lead to more detailed and fine-grained 3D scene reconstruction. However, as shown in Fig. 3, previous SoTA approaches (Wimbauer et al. 2023; Li et al. 2024) typically adopt a random patch sampler for uniform ray sampling across the entire scene, leading to redundant samples and overlooked instances. In contrast, our proposed SNOG sampler leverages informative visual priors from the pre-trained open-vocabulary model Grounded-SAM (Ren et al. 2024) (a combination of Grounding DINO (Liu et al. 2025) and SAM (Kirillov et al. 2023)) to optimize the allocation of computational resources while enhancing the awareness of crucial instances.

Specifically, we utilize the semantic labels from the Cityscapes dataset (Cordts et al. 2016) as prompts for Grounding DINO. After obtaining instance-level bounding boxes, we employ SAM to generate precise segmentation masks. Consequently, for the k-th instance, we acquire its metadata $\mathcal{M}_k = \{\boldsymbol{l}_k, \boldsymbol{b}_k, s_k\}$, where $\boldsymbol{l}_k$ denotes the center location of its bounding box, $\boldsymbol{b}_k$ stores half of the height and width of its bounding box, and s_k indicates the semantic area of the instance. Subsequently, we use Gaussian mixture distribution combined with background uniform distribution to achieve instance-aware and non-overlapping ray sampling, the probability density function (PDF) $p(\boldsymbol{x})$ of which can be formulated as follows:

$$p(\boldsymbol{x}) = (1-\gamma)\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) + \gamma \mathcal{U}\left(\boldsymbol{x} \mid s\right), \tag{4}$$

where $\mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ denotes the PDF of the bivariate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, $\mathcal{U}(\boldsymbol{x} \mid s)$ denotes the PDF of a 2D uniform distribution within the area s, and γ and π_k denote the weights of the background sampling and independent Gaussian distributions, respectively.

For the Gaussian distribution of the k-th instance, our objectives are to 1) locate $\boldsymbol{\mu}_k$ at the center of its bounding box and 2) ensure that approximately 95.5% of the samples fall within the bounding box. We can therefore initialize the parameters in (4) as follows:

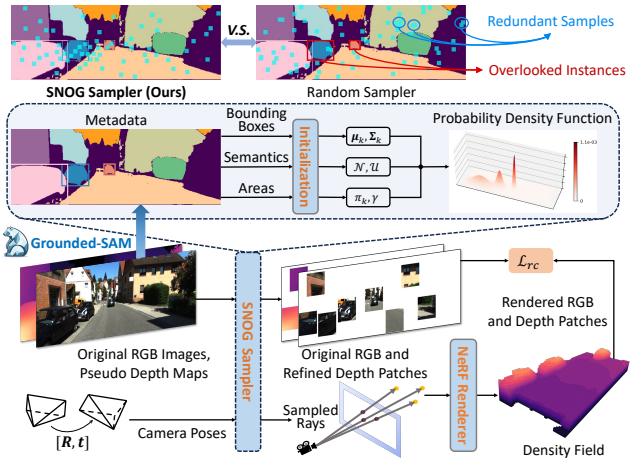

Figure 3: An illustration of our proposed SNOG sampler.

$$\begin{cases} \boldsymbol{\mu}_k = \boldsymbol{l}_k, \ \boldsymbol{\Sigma}_k = \operatorname{diag}\left(\frac{\boldsymbol{b}_k \circ \boldsymbol{b}_k}{4}\right) \\ \pi_k = \frac{\log s_k}{\log \prod_{k=1}^{K} s_k} \end{cases}, \text{ for } k = 1, ..., K, \tag{5}$$

where $\circ$ denotes the Hardmard product operation, and π_k is normalized in logarithmic space to prevent the sampling probability of smaller instances from approaching zero, especially when the semantic areas vary significantly among different instances.

Additionally, to address the redundant sampling issue, we incorporate constraints between the sampling PDF and existing samples, and formulate the final conditioned sampling PDF as follows:

$$P(\boldsymbol{x} \mid \mathcal{X}) = \begin{cases} 0, & \text{if } \exists\, \boldsymbol{x}_i \in \mathcal{X}, ||\boldsymbol{x} - \boldsymbol{x}_i||_2^2 < 2l^2 \\ p(\boldsymbol{x}), & \text{otherwise} \end{cases} \tag{6}$$

where l is the patch size, and $\mathcal{X}$ is an anchor set storing existing samples. With the final PDF, we randomly sample a collection of well-distributed and non-overlapping patches for image rendering and depth reconstruction. More details on the parameter initialization and the mathematical derivations of the PDF are given in our supplement.

Loss Formulation

A reference image can be warped to the target view using camera intrinsic parameters and differentiable grid sampling when its per-pixel depth is known. The original target image and the warped reference image should exhibit temporal photometric consistency. Furthermore, when performing volume rendering on a given frame, the rendered RGB and depth images should be respectively consistent with the original RGB image and the predicted metric depth map, thereby satisfying spatial reconstruction consistency. Therefore, we formulate a novel loss function as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ta} + \lambda_2 (\mathcal{L}_{rc}^{d} + \mathcal{L}_{rc}^{rgb}), \tag{7}$$

where $\mathcal{L}_{ta}$ denotes the temporal alignment loss, $\mathcal{L}_{rc}^{d}$ and $\mathcal{L}_{rc}^{rgb}$ represent the reconstruction consistency losses for

depth and RGB image rendering, respectively, and λ_1 and λ_2 are parameters used to balance these two types of losses.

Temporal Alignment Loss The homogeneous coordinates $\tilde{\boldsymbol{x}}^t$ and $\tilde{\boldsymbol{x}}^{t+1}$ in adjacent principal frames $\boldsymbol{I}_0^t$ and $\boldsymbol{I}_0^{t+1}$ are related as follows:

$$\tilde{\boldsymbol{x}}^{t+1} = \hat{\boldsymbol{D}}^t(\boldsymbol{x})\boldsymbol{K}\boldsymbol{T}\boldsymbol{K}^{-1}\tilde{\boldsymbol{x}}^t, \tag{8}$$

where $\boldsymbol{T}$ denotes the relative camera pose. Therefore, we can warp $\boldsymbol{I}_0^{t+1}$ into the pixel grid of $\boldsymbol{I}_0^t$ using differentiable grid sampling, producing a synthesized image $\hat{\boldsymbol{I}}_0^t$. The temporal alignment loss, expressed as follows:

$$\mathcal{L}_{ta} = \frac{1}{N}\sum_{\boldsymbol{x}} \boldsymbol{M}(\boldsymbol{x}) \left| \boldsymbol{I}_0^t(\boldsymbol{x}) - \hat{\boldsymbol{I}}_0^t(\boldsymbol{x}) \right|, \tag{9}$$

can be computed to enforce photometric similarity across adjacent frames, where $\boldsymbol{M}$ represents the weight mask detailed in (Bian et al. 2019) and N denotes the number of valid pixels for loss computation.

Reconstruction Consistency Loss It is common preliminaries that $\boldsymbol{D}(\boldsymbol{x})\tilde{\boldsymbol{x}} = \boldsymbol{K}\boldsymbol{p}$ and $||\boldsymbol{p}||_2 = \hat{\boldsymbol{D}}_r(\boldsymbol{x})$. We can therefore use these relations to establish criteria for depth reconstruction consistency as follows:

$$\mathcal{L}_{rc}^d = \frac{1}{M}\sum_{\boldsymbol{x}} \left| \frac{\hat{\boldsymbol{D}}_r(\boldsymbol{x})}{||\boldsymbol{K}^{-1}\tilde{\boldsymbol{x}}||_2} - \hat{\boldsymbol{D}}(\boldsymbol{x}) \right|, \tag{10}$$

where M denotes the number of valid pixels for loss computation. In addition, to enforce the consistency between the original and rendered RGB patches, we adopt the same rendering loss as detailed in (Wimbauer et al. 2023):

$$\mathcal{L}_{rc}^{rgb} = \beta_1 \,\mathrm{SSIM}\left(\boldsymbol{I}, \hat{\boldsymbol{I}}_r\right) + \beta_2 \left|\left| \boldsymbol{I} - \hat{\boldsymbol{I}}_r \right|\right|_1, \tag{11}$$

where $\beta_1 = 0.85$ and $\beta_2 = 0.15$ are the empirical parameters used in (Wimbauer et al. 2023).

Experiments

Datasets, Metrics, and Implementation Details

The 3D reconstruction performance of our proposed method is evaluated on the KITTI-360 dataset (Liao et al. 2022) and the KITTI Raw dataset (Geiger et al. 2013), both providing time-stamped stereo images along with ground-truth camera poses for the evaluation of scene perception algorithms. All images are resized to the resolution of 192×640 pixels, and the depth range is capped at 80m in both datasets. Following (Wimbauer et al. 2023), we split KITTI-360 dataset into a training set of 98,008 images, a validation set of 11,451 images, and a test set of 446 images for the 3D occupancy prediction task. We adopt the Eigen split (Godard et al. 2019) for depth estimation on the KITTI Raw dataset. Moreover, we use the DDAD dataset (Guizilini et al. 2020a) to evaluate our model's zero-shot generalizability using the weights obtained on the KITTI-360 dataset. The input images, with the original resolution of 1,216×1,936 pixels, are center-cropped and resized to 192×640 pixels for fair comparison.

Method	$\mathrm{O}^s_{acc}\uparrow$	$\mathrm{IE}^s_{acc}\uparrow$	$\mathrm{IE}^s_{rec}\uparrow$
Monodepth2 (Godard et al. 2019)	0.90	N/A	N/A
Monodepth2 + 4m	0.90	0.59	0.66
PixelNeRF (Yu et al. 2021)	0.89	0.62	0.60
BTS (Wimbauer et al. 2023)	0.92	0.69	0.64
KYN (Li et al. 2024)	0.92	0.70	0.66
ViPOcc (Ours)	**0.93**	**0.71**	**0.69**

Table 1: Comparison of scene reconstruction performance on the KITTI-360 dataset.

Method	$\mathrm{O}^o_{acc}\uparrow$	$\mathrm{IE}^o_{acc}\uparrow$	$\mathrm{IE}^o_{rec}\uparrow$
Monodepth2 (Godard et al. 2019)	0.69	N/A	N/A
Monodepth2 + 4m	0.70	0.53	0.52
PixelNeRF (Yu et al. 2021)	0.67	0.53	0.49
BTS (Wimbauer et al. 2023)	0.79	0.69	0.60
KYN (Li et al. 2024)	0.79	0.69	0.61
ViPOcc (Ours)	**0.79**	**0.69**	**0.64**

Table 2: Comparison of object reconstruction performance on the KITTI-360 dataset.

Following the experimental protocols established in previous works (Wimbauer et al. 2023; Li et al. 2024), we quantify the 3D occupancy prediction performance of the model using the following metrics: scene occupancy accuracy O^s_{acc}, invisible scene accuracy IE^s_{acc}, invisible scene recall IE^s_{rec}, object occupancy accuracy O^o_{acc}, invisible object accuracy IE^o_{acc}, and invisible object recall IE^o_{rec}. Furthermore, we use the mean absolute relative error (Abs Rel), mean squared relative error (Sq Rel), root mean squared error (RMSE), root mean squared log error (RMSE log), and accuracy under thresholds ($\delta_i < 1.25^i, i = 1, 2, 3$) to quantify the model's monocular depth estimation performance.

The proposed method is trained on an NVIDIA RTX 4090 GPU using the Adam optimizer for 25 epochs, with an initial learning rate of 1e-4, which is reduced by a factor of 10 during the final 10 epochs. We use BTS (Wimbauer et al. 2023) as our baseline network and adopt the metric depth predictions from Depth Anything V2 (Yang et al. 2024) as pseudo depth. We use Grounded-SAM (Ren et al. 2024) to generate instance-level semantic masks and bounding boxes.

Comparisons with State-of-The-Art Methods

3D Occupancy Prediction Following the experimental protocols detailed in the study (Wimbauer et al. 2023), we compare ViPOcc with a representative self-supervised monocular depth estimation network Monodepth2 (Godard et al. 2019) and other NeRF-based SoTA methods in terms of 3D occupancy prediction performance, as presented in Tables 1 and 2. Specifically, when evaluating Monodepth2's 3D occupancy prediction performance, all points behind visible pixels in the image are considered occupied. This is primarily due to the infeasibility of inferring the true 3D geometry of points that are invisible in the image. Furthermore,

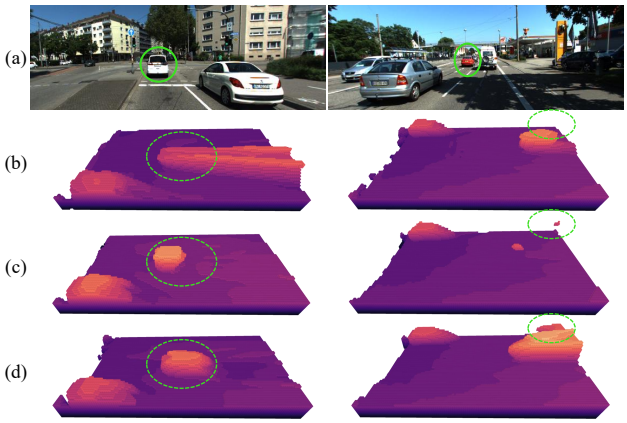

Figure 4: Qualitative comparison of 3D occupancy prediction on the KITTI-360 dataset: (a) input RGB images; (b) BTS results; (c) KYN results; (d) our results. A darker voxel color indicates a lower altitude.

Method	Abs Rel	RMSE log	$\delta < 1.25$
Pseudo depth (no scaling)	0.586	0.477	0.071
Pseudo depth (median scaling)	0.142	0.209	0.832
BTS (Wimbauer et al. 2023)	0.103	0.194	**0.891**
KYN (Li et al. 2024)	0.107	0.197	0.880
ViPOcc (Ours)	**0.097**	**0.188**	0.886

Table 3: Comparison of metric depth estimation performance on the KITTI-360 dataset.

we also follow prior studies (Wimbauer et al. 2023) to quantify the model's performance by considering points within a distance of up to 4m from visible points as occupied.

It can be observed that ViPOcc achieves SoTA performance across all metrics in 3D occupancy prediction for both scene and object reconstruction. Notably, O^s_{acc}, O^s_{rec}, and O^o_{rec} increase by 1.1-3.4%, 4.6-15.0%, and 4.9-30.6%, respectively. It is also worth noting that Monodepth2 + 4m can deliver competitive performance in O^s_{acc}. However, it relies on hand-crafted criteria rather than directly learning the 3D structure from a single view (Li et al. 2024).

Qualitative comparisons are presented in Fig. 4, where the predicted occupancy grids are viewed from the right side of the scene. It is evident that our method significantly outperforms both BTS and KYN in 3D geometry reconstruction, particularly for crucial instances, and effectively reduces trailing effects. These results demonstrate the efficacy of ViPOcc in reasoning about occluded regions against inherent ambiguities.

Metric Depth Estimation Table 3 shows the comparison of metric depth estimation performance among VFM, previous SoTA methods, and our proposed ViPOcc on the KITTI-360 dataset. It is evident that the depth predictions from Depth Anything V2 are unsatisfactory, regardless of whether median scaling is used to align the depth distribution. Notably, ViPOcc demonstrates superior performance compared

Method	Abs Rel	Sq Rel	$\delta < 1.25$
Monodepth2 (Godard et al. 2019)	0.106	0.818	0.874
SwinDepth (Shim and Kim 2023)	0.106	0.739	0.890
Lite-Mono (Zhang et al. 2023b)	0.107	0.765	0.886
BTS (Wimbauer et al. 2023)	0.102	0.755	0.882
MVBTS (Han et al. 2024)	0.105	0.757	0.873
KDBTS (Han et al. 2024)	0.105	0.761	0.873
ViPOcc (Ours)	**0.096**	**0.652**	**0.890**

Table 4: Comparison of depth estimation performance on the KITTI Raw dataset using the Eigen split.

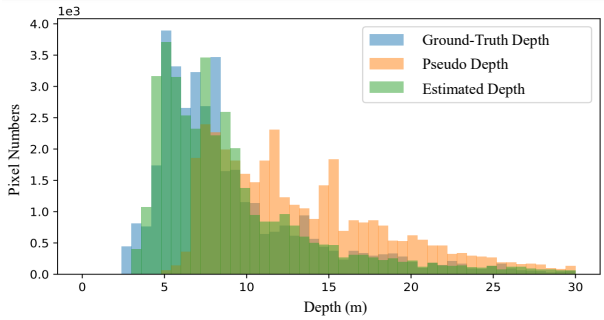

Figure 5: Depth distribution comparison.

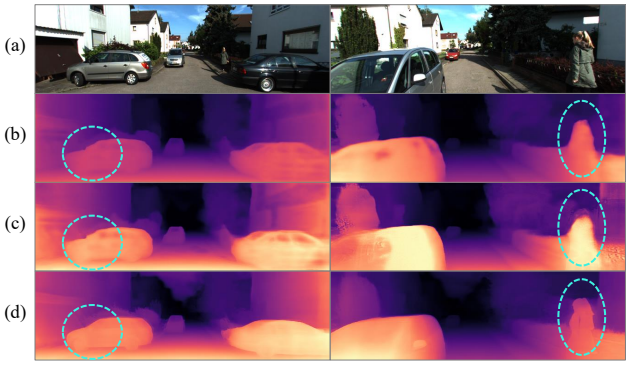

Figure 6: Comparison of metric depth estimation on the KITTI-360 dataset: (a) input RGB images; (b) BTS results; (c) KYN results; (d) our results.

to existing NeRF-based unsupervised methods, achieving a decrease of 5.8% in Abs Rel. Moreover, we compare depth distributions among ground-truth depth, pseudo depth, and our predictions. As illustrated in Fig. 5, significant discrepancies among these distributions not only underscore the infeasibility of directly using pseudo depth to generate supervisory signals, but also validate the effectiveness of our proposed inverse depth alignment module in refining depth.

As presented in Fig. 6, the qualitative comparison with prior SoTA methods on the KITTI-360 dataset demonstrates ViPOcc's exceptional metric depth estimation performance. Our method exhibits superior depth consistency in continuous regions, as shown on the vehicle's glass, and preserves

Configuration		O^s_{acc} ↑	IE^s_{acc} ↑	IE^s_{rec} ↑
Baseline (Wimbauer et al. 2023)		0.91	0.65	0.64
Depth estimation	+ Depth	0.91	0.64	0.64
	+ Pseudo depth	0.86	0.60	0.61
	+ Inverse pseudo depth	0.92	0.65	0.66
Ray sampling	+ Random sampler	0.91	0.65	0.64
	+ SNOG sampler	0.92	0.67	0.67
Loss	+ $\mathcal{L}_{ta}$	0.91	0.66	0.64
	+ $\mathcal{L}_{rc}$	0.89	0.64	0.60
	+ $\mathcal{L}_{ta}$, $\mathcal{L}_{rc}$	0.92	0.69	0.67
Full implementation		**0.93**	**0.71**	**0.69**

Table 5: Ablation studies of ViPOcc inner designs on the KITTI-360 dataset.

clear object boundaries, as shown on the pedestrian. These improvements can be attributed to the spatial reconstruction consistency constraint we incorporated between rendered and predicted depth maps, which also preserves the local differential properties of VFM predictions to enable fine-grained depth estimation.

Moreover, as presented in Table 4, ViPOcc also demonstrates superior depth estimation performance compared to all existing self-supervised methods on the KITTI Raw dataset. Specifically, it achieves a decrease of 5.9% in Abs Rel and 11.8% in Sq Rel compared to previous SoTA approaches. Surprisingly, ViPOcc significantly outperforms its counterparts trained with the same NeRF-based architectures, such as BTS (Wimbauer et al. 2023), MVBTS (Han et al. 2024), and KDBTS (Han et al. 2024). It achieves an average error reduction of 9.5% in Abs Rel and an average performance gain of 1.3% in δ_1. These experimental results underscore the effectiveness of our proposed ViPOcc framework for monocular depth estimation across different scenarios with distinct experimental setups.

Ablation Studies

We validate the rationality and efficacy of ViPOcc through extensive ablation studies, specifically focusing on depth estimation methods, ray sampling strategies, and loss function designs, as presented in Table 5.

We first adopt an individual depth prediction branch without VFM's visual priors incorporated for depth estimation, resulting in performance similar to that of the baseline. We attribute this phenomenon to a performance bottleneck within the depth prediction network, due to its estimations not being sufficiently accurate, which in turn limits improvements in 3D occupancy prediction. We then investigate the effectiveness of aligning VFM's depth priors based on depth residuals. As discussed earlier, neural networks struggle to converge or maintain accuracy when fitting depth residuals, which typically exhibit a large range of numerical variations. Consequently, a drastic performance drop occurs, falling within our expectations. When employing our proposed inverse depth alignment module, a notable performance improvement in 3D occupancy prediction is achieved, demonstrating its effectiveness.

Moreover, as observed, the SNOG sampler leads to improved performance, particularly in invisible scene accuracy and recall, which increase by 3.1-4.7%. This validates the effectiveness of our proposed ray sampling strategy. Additional comparisons between random and SNOG samplers regarding their efficiency are provided in our supplement.

In addition, it is evident that relying solely on temporal alignment loss yields limited performance improvements, whereas using only the reconstruction consistency loss actually degrades the framework's performance. However, combining both losses significantly enhances 3D occupancy prediction performance, leading to an increase of approximately 4.7-6.2% in invisible scene accuracy and recall.

Method	Abs Rel	RMSE log	$\delta < 1.25$
BTS (Wimbauer et al. 2023)	0.182	0.290	0.746
KYN (Li et al. 2024)	0.190	0.286	0.749
ViPOcc (Ours)	**0.175**	**0.282**	**0.749**

Table 6: Zero-shot depth estimation performance comparison on the DDAD dataset.

Zero-Shot Depth Estimation

To further evaluate the generalizability of ViPOcc, we conduct a zero-shot test on the DDAD dataset (Guizilini et al. 2020a) using the pre-trained weights obtained from the KITTI-360 dataset. As presented in Table 6, ViPOcc outperforms other SoTA methods in zero-shot depth estimation, demonstrating its exceptional generalizability.

Conclusion

This paper introduced ViPOcc, a novel framework that effectively leverages VFM's visual priors for single-view 3D occupancy prediction. ViPOcc consists of two coupled branches: one estimates highly accurate metric depth by aligning the inverse depth output from Depth Anything V2, while the other one predicts 3D occupancy with a Grounded-SAM-guided Gaussian mixture sampler incorporated to achieve efficient and instance-aware ray sampling. These two branches are effectively coupled through a temporal photometric alignment loss and a spatial geometric consistency loss. Extensive experiments and comprehensive analyses validate the effectiveness of our novel contributions and the superior performance of ViPOcc compared to previous SoTA methods. In the future, we aim to achieve a tighter coupling of these two branches and develop a more lightweight 3D occupancy prediction framework.

Acknowledgments

This project was supported by the National Natural Science Foundation of China under Grants 62473288, 62233013, and 62206046, the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (No. HMHAI-202406), the Fundamental Research Funds for the Central Universities, and Xiaomi Young Talents Program.

References

Bian, J.; et al. 2019. Unsupervised Scale-Consistent Depth and Ego-Motion Learning from Monocular Video. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Cao, A.-Q.; and De Charette, R. 2022. MonoScene: Monocular 3D Semantic Scene Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3991–4001.

Chen, X.; et al. 2023. Self-Supervised Monocular Depth Estimation: Solving the Edge-Fattening Problem. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5776–5786.

Cordts, M.; et al. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223.

Geiger, A.; et al. 2013. Vision Meets Robotics: the KITTI Dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Godard, C.; et al. 2019. Digging into Self-Supervised Monocular Depth Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3828–3838.

Guizilini; et al. 2020a. 3D Packing for Self-Supervised Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2485–2494.

Guizilini, V.; et al. 2020b. Semantically-Guided Representation Learning for Self-Supervised Monocular Depth. *arXiv preprint arXiv:2002.12319*.

Han, K.; et al. 2024. Boosting Self-Supervision for Single-View Scene Completion via Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9837–9847.

He, K.; et al. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, Y.; et al. 2023. Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9223–9232.

Huang, Y.; et al. 2024. SelfOcc: Self-Supervised Vision-Based 3D Occupancy Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19946–19956.

Jung, H.; et al. 2021. Fine-Grained Semantics-Aware Representation Enhancement for Self-Supervised Monocular Depth Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 12642–12652.

Kerr, J.; et al. 2023. LERF: Language Embedded Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 19729–19739.

Kirillov, A.; et al. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

L. Yang *et al.* 2024. Depth Anything V2. *arXiv preprint arXiv:2406.09414*.

Li, R.; et al. 2024. Know Your Neighbors: Improving Single-View Reconstruction via Spatial Vision-Language Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9848–9858.

Liao, Y.; et al. 2022. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3292–3310.

Liu, S.; et al. 2025. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *European Conference on Computer Vision*, 38–55. Springer.

Peng, S.; et al. 2023. OpenScene: 3D Scene Understanding with Open Vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–824.

Ren, T.; et al. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *arXiv preprint arXiv:2401.14159*.

Schmied, A.; et al. 2023. R3D3: Dense 3D Reconstruction of Dynamic Scenes from Multiple Cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3216–3226.

Shim, D.; and Kim, H. J. 2023. SwinDepth: Unsupervised Depth Estimation Using Monocular Sequences via Swin Transformer and Densely Cascaded Network. In *International Conference on Robotics and Automation (ICRA)*, 4983–4990. IEEE.

Sun, L.; et al. 2024. SC-DepthV3: Robust Self-Supervised Monocular Depth Estimation for Dynamic Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1): 497–508.

Sun, Y.; and Hariharan, B. 2024. Dynamo-Depth: Fixing Unsupervised Depth Estimation for Dynamical Scenes. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Tian, X.; et al. 2024. Occ3D: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Wang, W.; et al. 2023. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14408–14419.

Wang, Y.; et al. 2024. SQLdepth: Generalizable Self-Supervised Fine-Structured Monocular Depth Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 5713–5721.

Watson, J.; et al. 2021. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1164–1174.

Wei, Y.; et al. 2023. SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21729–21740.

Wimbauer, F.; et al. 2023. Behind the Scenes: Density Fields for Single View Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9076–9086.

Yang, L.; et al. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10371–10381.

Yin, W.; et al. 2021. Learning to Recover 3D Scene Shape from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 204–213.

Yin, Z.; and Shi, J. 2018. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1983–1992.

Yu, A.; et al. 2021. PixelNeRF: Neural Radiance Fields from One or Few Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4578–4587.

Zhang, C.; et al. 2023a. OccNeRF: Self-Supervised Multi-Camera Occupancy Prediction with Neural Radiance Fields. *arXiv preprint arXiv:2312.09243*.

Zhang, N.; et al. 2023b. Lite-Mono: a Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18537–18546.

Zhang, Y.; et al. 2024. Vision-Based 3D Occupancy Prediction in Autonomous Driving: A Review and Outlook. *arXiv preprint arXiv:2405.02595*.

Zheng, Y.; et al. 2024. MonoOcc: Digging into Monocular Semantic Occupancy Prediction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 18398–18405.

Zhou, T.; et al. 2017. Unsupervised Learning of Depth and Ego-Motion from Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1851–1858.